人工智能之拟人化*

喻丰 许丽颖
西安交通大学人文社会科学学院，西安 710049

摘　要：人工智能及其载体或者存在形式（如智能机器人、自动驾驶汽车等）不只具有智能，更具有社会功能。社会功能之具备得益于人工智能的拟人化。拟人化是将人类特征、动机、意向或心理状态赋予非人对象的心理过程或者个体差异，拟人化的产生受到激发主体知识、效能动机和社会动机的影响。对人工智能的拟人化影响存在如何拟人化（即拟人化的人工智能知觉以及破除恐怖谷效应等问题）、何时拟人化（即何种人格者在何种情境下更易拟人化人工智能）、拟人化何用（即拟人化人工智能的好处和意义）以及拟人化后的社会互动行为等问题。本文在心理学和人—机器人交互研究的基础上，讨论人工智能拟人化的如何、何时及为何问题，并对人工智能设计提供心理学思考。

关键词：人工智能，拟人化，社会机器人，人-机器人交互

1 何谓拟人化

拟人化通常是指一种将人类独有特质赋予非人实体的倾向性或形态（Epley, Waytz, & Cacioppo, 2007; 许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓, 2017）。这种倾向性或形态广见于人们的日常生活且颇具影响。它并非某些人的专利，上至老人，下至孩童，都有可能在有意或无意间将外物拟人化；它也不局限于单一的客体，上至神明，下至尘埃都有被拟人化的可能。虽然有学者（e.g., Caporael, 1986; Mitchell, 2005）将拟人化视为一种消极的认知偏差和不成熟的表现，但大量研究和实践都表明拟人化能够产生积极的后果（e.g., Tam, Lee, & Chao, 2013; Butterfield, Hill & Lord, 2012; de Visser et al., 2016）。如对自然的拟人化能够提升人们的环保意向（Tam et al., 2013）、对动物的拟人化能够促进动物保护（Butterfield et al., 2012）等。

基于其普遍性和可能产生的积极效应，拟人化经常被用于产品的设计和营销之中，深入人心的可口可乐“曼妙曲线”和 M&M 豆拟人化形象都是其成功应用的经典之作。随着科技的迅猛发展，人工智能已然成为拟人化应用的热门领域。科幻电影中的机器人往往都十分“像人”，“他/她们”拥有人类的外表、表现出人类的行为、甚至能够让人知觉到人类的思维、情绪和情感。由于机器人在现实生活中并不常见，因此普通大众对机器人的印象往往来源于这些科幻电影（Broadbent et al., 2010）。近 20 年来，社会机器人（social robots）的不断涌现使这些幻想逐步走入现实（Broadbent, 2017）。之所以称其为社会机器人，是因为它们的工作场景不再是与世隔绝的生产流水线，而是延伸到了教育（如早教）、商业（如购物向导）和医疗保健（如陪伴老人）等复杂的社会环境；它们的工作内容也不再只是重复单一的机械运动，而是扩展到了需要与人互动交流的社会活动。要在这些领域发挥作用，就要求社会机器人不能只是“工具”，而要成为人们的“社会伙伴”（Dumouchel & Damiano, 2017）。毫无疑问，在这个机器人由单纯的“机器”向真正的机器“人”逐渐蜕变的过程中，拟人化

* 国家自然科学基金青年项目（71501105）。通讯作者：喻丰，男，1985 年生于湖北武汉，现为西安交通大学社会心理学研究所教授、博士生导师，研究方向为道德心理学，E-mail: yufengx@xjtu.edu.cn。许丽颖，女，1993 年出生于湖北襄阳，现为西安交通大学与香港城市大学联合培养心理学博士研究生，研究方向为道德心理学。

能够起到强大的助推器作用。在心理的层面我们甚至可以说，正是由于拟人化的融入，使得机器人开始真正脱离简单的工具性，而更多地拥有复杂的社会性。

本文聚焦于人工智能尤其是社会机器人领域的拟人化问题，在心理学和人—机器人交互研究的基础上探讨如何拟人化（即拟人化的人工智能知觉以及破除恐怖谷效应等问题）、何时拟人化（即何种人格者在何种情境下更易拟人化人工智能）、拟人化何用（即拟人化人工智能的好处和意义）以及拟人化后的社会互动行为等问题。通过对这些话题的讨论，我们期望为人工智能设计提供有益的心理学建议，并以此引发更多对于人工智能拟人化的思考。

2 如何拟人化

人工智能实际上是基于代码的，如果剥茧抽丝来看人工智能，其底层应该是算法、代码与计算过程。但是人类需要具体化，算法相对抽象，而在许多时候我们需要一个具体而心理距离近的形象来表达这些算法（Trope & Liberman, 2010）。这便需要让人工智能拟人化。实际上，在民众心理学水平上，人们很难去理解或者解释为何底层代码能够运行出看似千变万化、无所不能的人工智能形态，抑或如果人工智能经由算法而没有表现出人的形态，实际上人们无法将其作为智能来看待。譬如日常生活中我们在街上随处可见所谓的机器人拉面，这种拉面机器实际上并非人工智能，而只是普通机器，但是由于其人形形象的存在，它通常情况下会被认为是人工智能。而真实的人工智能做饭或者炒菜的机器，由于其形态类似机器，而并无人形特征，在生活中常人却不会将其看做人工智能，知识匮乏者甚至无法理解这是人工智能的实体存在。甚至普通人开始制造其所谓的人工智能时，均会选择拟人化的形态而开始（Broadbent, 2017）。因此，似乎拟人化通常情况下可能成为理解和知觉人工智能是否时人工智能的窗口。

但实际上拟人化也有不同的形态。最基本的区分是外表拟人化与行为拟人化。行为拟人化相对研究偏少，但即使当一颗小球做出规则运动时，人们也会倾向于将其归结为有一颗心灵。同理，一个外表并不拟人化的机器形态的人工智能实体如若展现出某种人类行为的规律性，那么其是否拟人化这还有待研究。但是，外表拟人化的人工智能实体是在日常生活中常见的。我们可能发现，外表拟人化也有不同的表现形态。如有的人工智能拟人化形态就极尽像人，比如某些作为性爱用途的人工智能实体；有的人工智能机器人展现出人的特征，但是却在常人看来绝非人类，只是“像人”而已，比如机器人 Nao；有的人工智能机器人展现出婴儿特征，表现为萌的感觉（许丽颖, 喻丰, 印刷中），比如已开始商业化的机器人 Paro；有的人工智能机器人展现出宠物形态，其外形如狗、海豹、恐龙等，如大量研发的所谓机器狗等；有的人工智能机器人还表现出卡通特征等。那么我们是根据什么来确定人工智能实体像人的呢？研究发现，使用构建的拟人化机器人数据库（ABOT, Anthropomorphic roBOT）进行各种拟人特征评定，通过因素分析发现，人工智能实体或者机器人的特征可以分为四个部分，即表面外观（睫毛、头发、皮肤、性别、鼻子、眉毛、服装）、身体部位（手、手臂、躯干、手指、腿）、面部特征（面部、眼睛、头部、嘴巴）以及机械运动（车轮、踏板/履带），而这些维度对于是否像人的预测中，表面外观与身体部位总是最为可靠的预测源，而面部特征与机械运动相对更不重要，在所有的特征中，躯干、性别与皮肤是解释力最强的三个预测指标（Phillips, Zhao, Ullman, & Malle, 2018）。

但是否人工智能越像人越好呢？这里存在一个所谓的恐怖谷（uncanny valley）效应，即如果人工智能实体长得过于接近人，人们反而会感受到不适甚至恐惧（Mori, 1970）。这一效应是否存在实际上存在争议（MacDorman & Chattopadhyay, 2016）。以现有研究来看，如果一个人工智能机器人在外貌和行为上都与人无异，那么这实际上无问题。但如果有一方面略有奇怪，那么恐怖谷效应便会出现。譬如若机器人的外貌极其类似人，而其行为却不类似人或者其触感冰凉或无皮肤感，这时候就会出现恐怖谷效应（Cabibihan, Joshi, Srinivasa, Chan,

& Muruganantham, 2015)；而如若机器人的行为类似人，但外貌却是机械形态，恐怖谷效应也会产生（Pinar, Thierry, Hiroshi, Jon, & Chris, 2012）。当然，如果外貌或者行为其中一个维度上存在人与非人的混合，那么恐怖谷也会出现。比如眼神空洞的机器人会比正常眼睛的机器人更容易产生恐怖谷效应（Broadbent, Kumar, Li, Stafford, Macdonald, & Wegner, 2013）。为何会出现恐怖谷效应，这一点也存在不同理论的争议。恐怖谷的主要解释在于人们会将过于类似人的机器人看作是尸体，并因此产生不适与恐惧。但是为何我们会对尸体产生不适，这也存在争议。一种解释是，这可能是由于成熟或者文化教养的影响，因为对婴幼儿的研究发现大于 9 岁的儿童才会觉得类似人类的机器人比机械形态的机器人奇怪，而年龄更小的孩子却不会，这可能与孩子是否判断机器人有心灵有关系（Brink, Gray, & Wellman, 2018）。9 岁同时也是皮亚杰所谓他律道德转向自律道德的时期，这或许与其心智知觉有关（Gray, Gray, & Wagner, 2007）。当然第二种解释是远端解释，即在进化上人们倾向于避免与罹患疾病和死去的身体进行接触以保证健康（Broadbent, 2017）。

虽然会出现恐怖谷效应，而且在人们直觉上拟人化的机器人越像人越好。但是如若机器人能够融入社会之中，它们究竟是何形态合适，这却取决于不同的社会情境和任务。事实上，纵观现有研究，基本发现是如果一个人工智能实体所进行的任务是纯粹非社会行为，比如举重机器人、扫地机器人等，那么它不必呈现出任何拟人化形态（Broadbent et al., 2012）。而如若机器人需要承担社会任务，那么其拟人化则要优于非拟人化。如陪伴型的机器人需要更加拟人化的皮肤或者毛绒式的外表（这类似 Harry Harlow 的恒河猴陪伴实验）以及更加类似人类而非机器的声音（Broadbent et al., 2012; Tamagawa, Watson, Kuo, Macdonald, & Broadbent, 2011）。同样，如果陪伴型的机器人起到的是类似宠物那样与人类形成依恋的工作，则其最好也表现出宠物的模样（Broadbent, 2017）。但是也不是所有执行社会情境任务的机器人都需要拟人，比如有研究同时发现，医疗机器人也许不拟人化比较好，因为这样患者不容易尴尬（Bryant, 2010）。

3 何时拟人化

正如我们在前文中所提到的，普通大众对机器人其实有一种拟人化期望，而这很大程度来源于许多将机器人高度拟人化的科幻电影、科幻小说等媒介（Broadbent et al., 2010）。在这些科幻片里，有的机器人外表几乎能以假乱真，有的机器人行为与常人无异，有的机器人有意识有思想、敢爱敢恨，甚至还意图统治世界。当然，如今的机器人技术还远远没有达到科幻片或小说中描绘的程度，但不可否认的是，这些深入人心的形象极大地影响了普通人对于机器人的拟人化倾向。那么，除了这个显而易见的社会因素之外，从心理学的角度来看，还有哪些个体或情境因素影响了我们对机器人的拟人化呢？

针对何时拟人化这一问题，拟人化研究领域一个较为全面的理论框架来自 Epley 等人（2007）。他们从认知和动机两方面出发，提出拟人化的决定因素有三：第一是诱发主体知识（elicited agent knowledge），即对于人类而言，关于自身的知识自然比关于非人的知识要丰富得多，因此在面对自己不熟悉的非人客体时，比较容易用更可及（accessible）的人类知识来对其进行拟人化解释；第二是效能动机（effectance motivation），即人类有理解、掌控外部环境并与之互动的需要，这种需要促进了拟人化；第三则是社会动机（sociality motivation），即人类有社会交往的需要，而如果这种需要在其他人类身上得不到满足时，人们就会转而通过拟人化来弥补人际关系的缺失（Epley et al., 2007）。同样，这个基础的理论框架也可以用来解释人工智能何时会被拟人化的问题，且已经取得了一定成功（e.g., Eyssel & Kuchenbrandt, 2011）。

首先，当机器人越像人时，我们越容易将其拟人化。当认识对象与我们自身具有感知相似性（perceived similarity）时，我们会更依赖关于自我的知识对其进行推断和理解。如有研

究用 48 个不同的机器人考察影响拟人化的因素，结果发现与人类面部差异较大的机器人相比，具有更多人类面部特征（如嘴巴、鼻子、眼睛等）的机器人会被更多地拟人化（DiSalvo et al., 2002）。还有研究者（Kiesler, Powers, Fussell, & Torrey, 2008）直接对比了具有类似机器人功能的软件（如 Siri）和具有人类外貌特点的实体机器人，结果发现人们不仅对外表更像人的实体机器人拟人化程度更高，而且会在行为上透露给它更少的个人信息，因为它看起来"更逼真"（Kiesler et al., 2008）。当然，这种与人的相似性并不局限于外表，行为甚至社会关系的相似性也能起到类似的作用，来自神经生物学的证据证明了这一点。Hoenen，Lübke 和 Pause（2016）发现，当看到一个清洁机器人被言语骚扰时，人们会产生更强烈的同情，并且人脑中的镜像神经元（mirror neurons）会被更多地激活。他们认为，这是由于看到机器人的社会交往（这里只是单一方向的）增加了人们对机器人社会能动性的感知，从而促进了对机器人的拟人化（Hoenen et al., 2016）。

其次，当我们自身的效能动机越强时，越容易将机器人拟人化。效能动机是指人们掌握外部环境并与之互动的需要（White, 1959）。面对不熟悉的客体时，对确定性和控制力的寻求会对拟人化产生促进作用，因为拟人化能够最方便快捷地消除不确定性，增加预测性。虽然人们对科幻电影和小说中的机器人并不陌生，但在实际生活中能够经常接触到机器人的人并不多。也就是说，我们大部分人对机器人其实并不熟悉，在真正面对这种未知的科技产品时，我们会自然而然地产生心理紧张和不确定性，且在机器人不可预测或不可控的情况下尤甚（Eyssel & Kuchenbrandt, 2011）。效能动机对拟人化的促进作用已经得到了一些研究的验证（e.g., Epley, Waytz, Akalis, & Cacioppo, 2008; Waytz et al., 2010），与之相关的闭合需求（need for closure）、控制欲（desire for control）和个人结构需求（personal need for structure）等因素也被认为是影响拟人化的个体差异变量。如研究发现有稳定控制需要的人更容易拟人化看起来不可预测的动物（Epley, et al., 2008），被描述为不可预测的电脑、科技产品也会被更多地拟人化（Waytz et al., 2010）。Eyssel 和 Kuchenbrandt（2011）直接对在实验中机器人 NAO（阿尔德巴兰机器人公司开发的一款类人机器人）进行考察，他们通过操纵其行为可预测的程度，在人工智能领域验证了效能动机对拟人化的影响（Eyssel & Kuchenbrandt, 2011）。

最后，当我们自身的社会动机越强时，越容易将机器人拟人化。保持与他人的社会联系是人的基本需求之一，一般来说，我们会通过与他人的交往来满足这项需求，但当这种途径无法满足社会需求时，我们也可以通过将非人物体拟人化来得到一定程度的弥补。研究发现，长期的孤独（e.g., Epley, Waytz et al., 2008）和依恋焦虑（e.g., Bartz et al., 2016）等都会增强人们对非人对象的拟人化。即使只是在实验中暂时启动被试的孤独感（Epley, Akalis, Waytz, & Cacioppo, 2008）或者通过游戏使被试感到被他人排斥（Chen, Wan, & Levy, 2017），也会对拟人化及其后续行为造成影响。在人工智能特别是机器人领域，拟人化的社会动机显得尤为明显。社会机器人的出现恰好迎合了人们的社会需求，其中拟人化对于促进社会机器人与人之间的互动、进而满足人们的社会需求起到了不容忽视的作用。已经有研究表明，孤独的人比不孤独的人更容易将类人机器人拟人化（Eyssel & Reich, 2013），实际生活中也已经有针对老年人社会需求的陪伴型社会机器人面世。日本的 Paro 机器人就是其中很著名的一款，它像一只可爱的海豹，能够对人们的声音和抚摸等做出反应。虽然它的外形并不像人，但由于它能做出一些类人的反应，因此很多主人会不自觉地将它拟人化，觉得它有意识、有情绪、有身体状态（如感到寒冷）（Broadbent, 2017），并通过与之互动满足自己的社会需求。研究发现，Paro 能够减少人们的孤独感（Robinson, MacDonald, Kerse, & Broadbent, 2013）甚至减轻痴呆患者的躁动和抑郁（Jøranson, Pedersen, Rokstad, & Ihlebæk, 2015）。

4 拟人化何用

何时拟人化是从使用者的角度来探究何种人格者在何种情境下更易拟人化人工智能，而从人工智能设计者的角度而言，为何要将人工智能拟人化也是值得考虑的一个话题。那么，除了满足使用者对于人工智能特别是机器人的拟人化期望以外，设计者拟人化人工智能还可能出于何种目的和理由呢？或者换言之，人工智能拟人化的好处或者意义又究竟何在？

当然拟人化有其理论意义，而这体现在心理学领域。这种理论意义并不是我们通常我们所说的为心理学研究提供理论指导，而是可以在一定程度上弥补心理学研究的局限（如以人为被试的相关伦理问题），对已有的心理学理论以新的方式验证和拓展之，对未知的领域以可行的方式探索之。

一方面，人工智能拟人化可以作为心理学研究的试验场。即用拟人化的人工智能尤其是机器人来作为实验的“材料”甚至被试，能够规避以往心理学研究中存在的诸多干扰因素和伦理问题（Jakub, Proudfoot, Yogeeswaran, & Christoph, 2015）。心理学是一门研究人类心理与行为的学科，而人类恰恰是最复杂的动物，虽然从理论上来说，心理学实验应当严格控制所有可能会影响研究结果的干扰变量，但是这在实际的操作中其实是很难实现的。举一个简单的例子，如果我们想要研究别人的欺骗行为对人们心理及后续行为的影响，我们可能会给被试播放一段我们预先录制好的欺骗行为的视频，而这样的操作显然就存在两个问题：一是录制这个视频的“演员”本身的个人特点（如外貌、表达方式等等）可能会对实验结果产生影响，二是这种实验并不能让被试产生身临其境的真实感，这也可能会使实验的效果大打折扣。而拟人化机器人恰恰能帮忙解决这两个问题，我们可以大胆想象，当机器人达到恰当的拟人化程度后，我们可以在实验中让被试直接观察其欺骗行为。它不仅能够通过诱发人们的拟人化倾向（即把机器人当成人）达到相似的实验效果，而且便于实验者控制实验中的各种干扰变量。此外，以往的许多经典心理学研究都因其伦理问题而备受诟病（如臭名昭著的服从实验和斯坦福监狱实验等），而以拟人化的机器人来代替实验中的“真人”，不仅能够在不触碰伦理红线的前提下重复验证这些实验结果，并且能够对比人们对真人和机器人的反应。迄今为止，已经有一些研究者（e.g., Bartneck, Rosalia, Menges, & Deckers, 2005; Hoffman et al., 2015）用这种方法对之前的研究进行了重复和检验，如 Hoffman 等人（2015）重复了 Covey，Saladin 和 Killen（1989）关于旁观者能够增加诚实的实验，结果发现机器人在场也同样能够增加诚实（Hoffman et al., 2015）。类似地，米尔格拉姆的服从实验（Milgram, 1963）也得到了重复，研究者用机器人代替真人演员接受电击，结果发现与原实验中 65%的被试会将电压加到最大伏特相比，机器人实验中所有被试都加到了最大伏特，这不仅再次验证了权力服从，也表明人们对待机器人的态度与对待人类有所不同（Bartneck et al., 2005）。

另一方面，人工智能拟人化可以作为心理学研究的模拟器。虽然我们可以通过外显行为推断人的心理，但我们始终无法直接观察到人的内心活动和发展过程，而人工智能拟人化或许可以通过模拟人心而成为打开这一“黑箱”的工具，帮助人们理解人类的思维过程和心理发展过程。“人工智能之父”图灵在回答“机器能否思考”这一问题时提出了著名的图灵测试，并指出要创造“儿童机器”（child machine）（Turing, 1950）。所谓“儿童机器”，就是指具有儿童学习能力的机器，而制造这种机器的关键就在于模拟儿童学习的过程（Proudfoot, 2015）。值得注意的是，图灵同时还强调了类人具身（human-like embodiment）的重要性（Turing, 1950），这与拟人化有着密不可分的联系。可以说，心理学和人工智能在这一点上达到了真正的互促互进：人工智能拟人化、制造这种人类级人工智能的过程，可以帮助人们更多地了解以前无法被观察的认知甚至情绪情感获得的过程；而更多地了解我们人类自身，无疑也能够推动人工智能不断向真正的人类级迈进。

当然，除开心理学上的理论意义，人工智能拟人化的实践意义主要体现在人—机器人交互领域。有学者（Disalvo & Gemperle, 2003）曾指出，对产品进行拟人化设计主要有四个作用：第一是保持事物一致性（keeping things the same），即有些产品长期以来就以拟人化形

象示人，拟人化能够保持其外形、风格的一贯性，避免不必要的疑惑和不适应；第二是解释未知（explaining the unknown），即拟人化设计能够帮助人们理解新产品的性能；第三是反应产品特性（reflecting product attributes），即用拟人化的外形凸显产品的质量、特色等；第四是展现人类价值（projecting human values），即通过拟人化设计传递个人、社会或文化价值（Disalvo & Gemperle, 2003）。这四点同样也可以适用于人工智能的拟人化，但总体而言，人工智能尤其是社会机器人的拟人化主要的积极作用主要在于促进人—机器人之间的良性互动（Damiano & Dumouchel, 2018）。

一方面，从使用者的体验来看，拟人化能够提升使用者对于人工智能的熟悉度和信任感。如前所述，人们拟人化人工智能的一个重要动机就是效能动机，面对较为陌生的人工智能，拟人化能够大大提升人们的亲切感，帮助使用者以最简单的方式理解其性能并迅速进行掌控。信任也是人工智能领域的热点问题，高科技的发展日新月异，但是要想真正走进千家万户，提升人们对人工智能的信任是一个重要的前提。研究发现，与非拟人化的无人驾驶汽车相比，拟人化的无人驾驶汽车会被给予更多的信任和宽容（Waytz, Heafner & Epley, 2014）；类似地，具有拟人化特征的机器人也会得到人们更多的信赖（Leite, Pereira, Mascarenhas, Martinho, Prada, & Paiva, 2013）。这些研究结果都表明了在人工智能领域，拟人化对提升信任感的有效作用。

另一方面，从设计和制造者的角度而言，拟人化有助于保护人工智能免受破坏并得到更多训练。人工智能，尤其是经常出现在公共场所为人们服务的社会机器人，经常会受到人们的破坏和侮辱（Rehm & Krogsager, 2013），这不仅会造成大量的经济损失，更被视为一种不人道的行为。为此，还有人自发组织了停止机器人虐待（stop robot abuse）的活动，呼吁人们停止虐待，关注机器人的各项权利。而相关研究表明，拟人化或许能够在这方面发挥重要影响。如有研究发现，与非拟人化机器人相比，人形机器人会得到更多的赞扬和更少的惩罚（Bartneck, Reichenbach, & Carpenter, 2006）。此外，就像我们在前文中所提到的，人工智能的一个重要目标在于创造“儿童机器”，这种机器需要通过与人的互动不断地学习和训练，而拟人化恰恰能够促进人机互动，从而使机器得到更多学习和训练的机会。

5 拟人化何解

既然拟人化的人工智能形态似乎是在民众水平知觉上以及研究者的科学研究中无可避免，那么拟人化存在的理论争议究竟该如何解决这是需要我们最终探索与回答的问题。

第一，拟人化的人工智能实体是否会影响社会关系？批评的声音认为，拟人化的机器人确实能够与人建立起社会联系，但是这种联系是否真的是社会的是存疑的（Damiano & Dumouchel, 2018）。因为拟人化的机器人毕竟不是人，这种人与非人的关系只是拟人化外表赋予人与算法之间建立的关系，究其实质，这种关系是虚伪和欺骗甚至自我欺骗的关系。长久沉浸的与拟人化机器人建立起的关系，这甚至会反而影响人类真实的社会关系。社会机器人的出现会打乱人际结构和社会秩序，甚至带来诸多伦理问题。比如陪伴型的拟人化机器人若与人类形成了稳定的联结，那么人类沉浸其中，还是否能够实践人类社会运行的规则与潜规则，类似中国人如此复杂的关系社会还是否需要学习和运用都值得思考，也许机器人只需要通过设置就能够完全如我们所愿，而不需要付出沉重而复杂的人类社会经验积累过程。这甚至会让潜移默化指导人类行为的文化规则失效，而真正消灭人类累计的文化（喻丰, 彭凯平, 2018）。同理，如果拟人化的性爱机器人和人类能够形成稳定联系，这是否会增加人类的暴力、虐待行为，在某种程度上重塑人类亲密关系的同时，是否会影响人类生存与繁衍都是值得深思的问题。拟人化的人工智能实体确实会对社会关系造成影响，但这在于人类将这些人工智能用于何处以及如何使用，也许对其规范性的应用能够解决此问题。

第二，拟人化的人工智能实体是否会威胁人类？人工智能威胁人类是我们长期的假设与

忧虑，这种忧虑关乎人类自身。当人工智能还只是算法时，由于其解释水平高，因此人工智能算法对普通人感知到的威胁并不大。但如果这种算法一旦鲜活具体，比如拟人化的人工智能实体或者机器人活生生地出现在人类面前，那么我们无法不将其视作某种物种或者类似我们自己的生物体。这种非常像人的机器人不仅会被知觉为对人类工作、安全、资源的真实威胁，而且会被视为对人类身份和独特性的威胁，尤其是如果这种机器人还能胜过人类时（Yogeeswaran, Złotowski, Livingstone, Bartneck, Sumioka, & Ishiguro, 2016）。研究还发现，当人们被告知机器人能力比人类强时，拟人化成都更高的机器人会使得人们知觉到更大的威胁，人们甚至会因此对机器人研究的支持下降（Yogeeswaran et al., 2016）。某种程度上，是否拟人化对人工智能是否会威胁人类并无关系，这并非一个本体论问题，而只是存在于认识论上。亦即，这是一个人类知觉问题。从这个意义上来说，无论人工智能是否拟人化，其威胁真实存在且本质上不会受到影响，受到影响的是人类知觉威胁的程度不同。

第三，拟人化的人工智能实体是否适用于人类规范？一旦机器人成功拥有拟人化形态，那么我们便倾向于将其当做人类看待。比如让拟人化的机器人进行观点采择任务时，我们会倾向于采用机器人的第一人称视角来看待世界，而相对不太容易去使用心理理论看待非你人化的机器人，正如我们不倾向于用动物视角看待外部世界一样（Zhao, Cusimano, & Malle, 2016）。事实上，人类不倾向于让人工智能去做道德判断（Bigman & Gray, 2018），但如若人工智能今后渗入生活作为道德责任归因主体时，我们必须对其进行责任归因（Bonnefon, Shariff, & Rahwan, 2016; Shariff, Bonnefon, & Rahwan, 2017; Awad, Dsouza, Kim, Schulz, Henrich, Shariff, Bonnefon, & Rahwan, 2018）。对机器道德责任归因的研究发现，如果面临杀一救五的列车困境时（喻丰, 彭凯平, 韩婷婷, 柴方圆, 柏阳, 2011），救或不救即作为或不作为，对人来说，作为时人类会受到更多的道德责任归因，而不作为时，机器会受到更多的道德责任归因（Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015）。而当机器人拟人化时，人们倾向于在作为情况下，更多谴责拟人化的机器人；而在不作为情况下，更多谴责非拟人化的机器人（Malle, Scheutz, Forlizzi, & Voiklis, 2016）。与机器人进行交流时，人类自己会显示出与人交流的不同，如更加礼貌（Nass, Moon, & Carney, 1999）；但是讽刺的是，人类却会不自觉地将人类社会规范应用于拟人化的人工智能。比如研究发现人们评价机器人时也会存在内群体偏见，被试不仅对内群体机器人评价更好，而且也更多将其拟人化（Eyssel & Kuchenbrandt, 2012）。这两点应该都是值得商榷的。人类通常情况下以为自己在无偏不倚、不带歧视、公正平等、保障权益地在与拟人化人工智能进行社会交往，但是却又不自觉地自己表现出社交行为的不同，这本身便是矛盾。接着人类又将自己的社会规范甚至于道德自动化地赋予拟人化人工智能，这则更加一重矛盾。也许，我们不应该将社会规范与人类道德直接加诸于机器人，而应该去寻求发展一套独立于人类而适用于拟人化人工智能的规范体系。当然，实验伦理学（彭凯平, 喻丰, 柏阳, 2011）研究者也应该抓住此契机来深化道德研究。

参考文献

彭凯平, 喻丰, 柏阳. (2011). 实验伦理学：研究、挑战与贡献. *中国社会科学,* (6), 15–25.

许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓. (2017). 拟人化:从“它”到“他”. *心理科学进展, 25*(11), 1942–1954.

许丽颖, 喻丰. (印刷中). 萌:感知与后效. *心理科学进展.*

喻丰, 彭凯平, 韩婷婷, 柴方圆, 柏阳. (2011). 道德困境之困境：情与理的辩争. *心理科学进展, 19*(11), 1702–1712.

喻丰, 彭凯平. (2018). 文化从何而来. *科学通报, 63*(1), 32–37.

Bartneck, C., Rosalia, C., Menges, R., & Deckers, I. (2005, September). Robot abuse—a limitation of the media equation. *Proc. Interact 2005 Workshop Abuse* (pp. 54–57)*,* Rome.

Bartneck, C., Reichenbach, J., & Carpenter, J. (2006, September). Use of praise and punishment in human-robot collaborative teams. *The IEEE International Symposium on Robot and Human Interactive Communication* (pp.177–182), Hatfield, UK.

Bigman, Y. & Gray, K. (in press). People are averse to machines making moral decisions. *Cognition*.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Brink, K. A., Gray, K., & Wellman, H. M. (in press). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development*.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*(1), 627–652.

Broadbent, E., Kumar, V., Li, X., Stafford, R. Q., Macdonald, B. A., & Wegner, D. M. (2013). Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality. *Plos One, 8*(8), e72589.

Broadbent, E., Kuo, I. H., Lee, Y. I., Rabindran, J., Kerse, N., & Stafford, R., et al. (2010). Attitudes and reactions to a healthcare robot. *Telemedicine and e-Health, 16*(5), 608–613.

Broadbent, E., Tamagawa, R., Patience, A., Knock, B., Kerse, N., & Day, K., et al. (2012). Attitudes towards health-care robots in a retirement village. *Australasian Journal on Ageing, 31*(2), 115–120.

Bryant, T. (2010). The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn, 1*(2), 109–115.

Butterfield, M. E., Hill, S. E., & Lord, C. G. (2012). Mangy mutt or furry friend? Anthropomorphism promotes animal welfare. *Journal of Experimental Social Psychology, 48*, 957–960.

Cabibihan, J. J., Joshi, D., Srinivasa, Y. M., Chan, M. A., & Muruganantham, A. (2015). Illusory sense of human touch from a warm and soft artificial hand. *IEEE Transactions on Neural Systems & Rehabilitation Engineering, 23*(3), 517–527.

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior, 2*(3), 215–234.

Chen, R. P., Wan, E. W., & Levy, E. (2017). The effect of social exclusion on consumer preference for anthropomorphized brands. *Journal of Consumer Psychology, 27*, 23–34.

Covey, M., Saladin, S., & Killen, P. (1989). Self-monitoring, surveillance, and incentive effects on cheating. *Journal of Social Psychology, 129*(5), 673–679.

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology, 9*, 468.

de Visser, E. J., Monfort, S. S., Mckendrick, R., Smith, M. A. B., Mcknight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*, 331–349.

Disalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002, June). All robots are not created equal: The design and perception of humanoid robot heads. (pp.321–326). DBLP. Proc. Des. Interact. Syst., 4th, London, Jun. 25–28, pp. 321–26. New York: ACM

Disalvo, C., & Gemperle, F. (2003). From seduction to fulfillment: The use of anthropomorphic form in design. *International Conference on Designing Pleasurable Products and Interfaces, 2003, Pittsburgh, Pa, Usa, June* (pp.67–72). DBLP.

Dumouchel, P., & Damiano, L. (2017). *Living with Robots*. Cambridge, MA: Harvard University Press.

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science, 19*, 114–120.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*, 864–886.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition, 26*, 143–155.

Eyssel, F., & Kuchenbrandt, D. (2011). Manipulating anthropomorphic inferences about NAO: The role of situational and dispositional aspects of effectance motivation. *Ro-man* (pp.467–472). IEEE.

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robotgroup membership. *British Journal of Social Psychology,51*(4), 724–731.

Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots): On the effects of loneliness on psychological anthropomorphism. In H. Kuzuoka (Ed.), *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 121–122). Piscataway, NJ: IEEE Press.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315,* 619.

Hoenen, M., Lübke, K. T., & Pause, B. M. (2016). Non-anthropomorphic robots as social entities on a neurophysiological level. *Computers in Human Behavior, 57*, 182–186.

Hoffman G, Forlizzi J, Ayal S, Steinfeld A, Antanitis J, et al. 2015. Robot presence and human honesty: experimental evidence. Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact., 10th, Portland, OR, Mar. 2–5, pp. 181–88. New York: ACM

Jakub, Z., Proudfoot, D., Yogeeswaran, K., & Christoph, B. (2015). Anthropomorphism: opportunities and challenges in human–robot interaction. *International Journal of Social Robotics, 7*(3), 347–360.

Jøranson, N., Pedersen, I., Rokstad, A. M., & Ihlebæk, C. (2015). Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: A cluster-randomized controlled trial. *Journal of the American Medical Directors Association, 16*(10), 867–873.

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26*, 169–181.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International Journal of Human - Computer Studies, 71*(3), 250–260.

Macdorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition, 146*, 190–205.

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371–378.

Mitchell, S. D. (2005). Anthropomorphism and cross-species modeling. In L. Daston & G. Mitman (Eds.), *Thinking with Animals* (pp. 100–118). New York: Columbia University Press.

Mori, M. (1970). The uncanny valley. *Energy, 7,* 33–35.

Nass, C., Moon, Y., & Carney, P. (1999). Are respondents polite to computers? Social desirability and direct

responses to computers. *Journal of Applied Social Psychology, 29*(5), 1093–1110.

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is human-like?: Decomposing robot human-like appearance using the Anthropomorphic roBOT (ABOT) Database. In HRI '18: Proceedings of the Eleventh Annual ACM/IEEE International Conference on Human-Robot Interaction, Chicago, Illinois, USA. Piscataway, NJ: IEEE Press.

Pinar, S. A., Thierry, C., Hiroshi, I., Jon, D., & Chris, F. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience, 7*(4), 413–422.

Proudfoot, D. (2015) Turing's child-machines. In Bowen J, Copeland J, Sprevak M, & Wilson R (Eds), *The turing guide: Life, work, legacy.* Oxford: Oxford University Press.

Rehm, M., & Krogsager, A. (2013). Negative affect in human robot interaction–impoliteness in unexpected encounters with robots. *Ro-Man* (Vol.54, pp.45–50). IEEE.

Robinson, H., Macdonald, B., Kerse, N., & Broadbent, E. (2013). The psychosocial effects of a companion robot: A randomized controlled trial. *Journal of the American Medical Directors Association, 14*(9), 661–667.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694–696.

Tam, K. P., Lee, S. L., & Chao, M. M. (2013). Saving Mr. Nature: Anthropomorphism enhances connectedness to and protectiveness toward nature. *Journal of Experimental Social Psychology, 49*, 514–521.

Tamagawa, R., Watson, C. I., Kuo, I. H., Macdonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics, 3*(3), 253–262.

Trope Y, & Liberman N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52,* 113–117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology, 99*, 410–435.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297–333.

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction, 5,* 29–47.

Zhao, X., Cusimano, C., & Malle, B. F. (2016). Do people spontaneously take a robot's visual perspective? *Proceedings of the 2016 ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*.

Artificial Intelligence and Anthropomorphism

YU Feng XU Liying

School of Humanities and Social Sciences, Xi'an Jiaotong University, Xi'an 710049, China

Abstract: Artificial intelligence (AI) and its forms (e.g., intelligent robots, autonomous cars) have both intelligence and social functions, and the social functions benefit from anthropomorphism of artificial intelligence. Anthropomorphism refers to the psychological process or individual difference of imbuing nonhuman agents with humanlike characteristics, motivations, intentions, or mental states, and it is influenced by elicited agent knowledge, effectance motivation, and sociality motivation. There are several questions worthy of consideration in the field of AI and anthropomorphism. First, how do we anthropomorphize AI? Second, when will we anthropomorphize AI? Third, what are the advantages of anthropomorphizing AI? Fourth, what else should we pay special attention to after anthropomorphizing AI? Based on the research in psychology and human-robot interaction, this paper discusses how, when and why we should anthropomorphize AI, and hoping to provide psychological references for AI design.

Key words: artificial intelligence; anthropomorphism; social robot; human-robot interaction